图书情报工作
第64卷 第18期 2020年9月

社会化问答社区答题者发现研究*

■ 潘梦雅 沈旺 代旺 刘嘉宇
吉林大学管理学院 长春 130022

摘 要：[目的/意义]识别社会化问答社区中回答可能性高的专业答题者，可缩短提问用户得到满意答案的等待时间，促进用户间的知识共享，助力社会化问答社区的持续健康发展。[方法/过程]基于社会资本理论及动机理论，对用户答题动因进行分析，结合专家发现研究提出测量指标，构建研究模型，以知乎社区为研究实例，借助 Python 语言对实验数据进行特征值提取、打标签等数据处理，研究运用逻辑回归模型、随机森林、XGBoost3 种常用的机器学习分类模型进行训练及预测。[结果/结论]与 PageRank、HITS 算法对比验证本文方法的有效性及优越性，本研究为同类平台如健康社区的问题推送、专家识别以及推荐模型的课题研究提供一定的参考。

关键词：社会化问答社区 专家发现 社会资本理论 动机理论 机器学习
分类号：G202 G206
DOI：10.13266/j.issn.0252-3116.2020.18.009

1 引言

互联网技术的发展，改变着人们搜寻和交流信息的方式，也带来了网络问答社区的兴起与繁荣。这些网络问答社区跨越时空限制，整合了不同背景、不同行业中具有相同或相似兴趣、目标和实践经历的用户群体，突破了仅仅通过搜索引擎搜索互联网上已有单一信息的信息获取约束，将用户大脑中的信息、经验、知识转移到网络问答社区。用户可以随时提出问题或回答任何领域和不同类型的问题，或通过评论、私信的方式与社区其他用户进行某种程度的即时交流，共享经验和知识，解决实际问题。

但在志愿式参与的网络问答社区中，仍然存在用户提出的问题长时间得不到回应，或问题得不到专业性、完整性、满意度较高答案的现象。久而久之，提问者会产生沮丧情绪，并可能影响社区的整体健康[1]。因此，识别社会化问答社区中针对特定问题有较高回答几率的专业答题者，能够使提问者得到高质量的回答，缩短用户得到满意答案的等待时间，促进社区的持续健康发展。此前也有学者探讨了如何识别问答社区某话题领域内的专家用户，实际上，若专家用户受各种条件限制，无法及时回答问题时，社区中问题得不到回复或得不到满意回复的现状仍难以改善。因此，本文借助动机理论和社会资本理论，结合专家发现的相关研究，力求找到专业且具备较大答题可能性的回答者，以解决以上问题。本文采用多种方法验证研究模型，以找出本模型的最优算法，不同于以往单一算法支撑下的研究，实验结果验证本研究的有效性及优越性。

本研究以热门社会化问答社区知乎为例（截至2019年1月，据 ALEXA 排名显示，"知乎"居我国社交网站排名第3位，居全球网站排名第90位，日均 IP 访问量约500万），从知乎的医学话题领域抓取研究数据，因为医学话题是一个专家用户及普通用户均可广泛参与的话题，故研究样本具有代表性。

2 相关研究综述

本研究的目的是寻找社会化问答社区答题可能性大的专业答题者，因此本节回顾专家发现及用户知识共享的相关研究。

2.1 专家发现

社会化问答社区专家发现就是从众多的回答者中寻找出掌握专业知识且权威可信的用户[2]，此前已有众多学者采用不同的研究方法，从不同的角度对问答社区的专家发现进行了研究探讨：

* 本文系国家自然科学基金项目"基于图模型的多源异构在线产品评论数据融合与知识发现研究"（项目编号：71974075）研究成果之一。
作者简介：潘梦雅（ORCID：0000-0002-0319-626X），硕士研究生，E-mail：pmy156@126.com；沈旺（ORCID：0000-0002-8933-5653），副教授；代旺（ORCID：0000-0001-7168-7776），硕士研究生；刘嘉宇（ORCID：0000-0002-2317-8157），硕士研究生。
收稿日期：2020-04-26 修回日期：2020-06-16 本文起止页码：76-88 本文责任编辑：王传清

(1)从问答社区的内容主题角度开展研究：J. Weng 等[3]依据用户的推特分布、推特内容的同质性，采用 TwitterRank 算法进行了主题敏感度的推特专家排名；A. Pal 等[4]通过话题内容的聚类，借助高斯混合模型，根据用户特性识别特定话题的权威专家；Z. Yan 等[5]利用张量模型和主题模型，研究了问题和回答者之间的潜在语义关系，通过 AUC 的最大化实现对潜在回答者的排名。

(2)从问答社区用户反馈行为来识别专家用户：X. Cheng 等[6]依据用户反馈作为相关标签词并建立主题模型，结合用户专业知识特征排序，最终实现专家发现；J. Shen 等[7]基于用户点赞、评论、选择最佳答案等用户反馈行为，通过加权的 HITS 算法推荐专家；S. Patil 等[8]分析了专家与非专家的行为，基于用户活动特征、答案质量特征、语言特征和时间特征 4 个指标，使用统计模型发现专家。

(3)以问答社区用户间的相似性或用户的社交网络关系为出发点，进行专家发现，如龚凯乐等[9]基于“问题-用户”的传播网络，拓展用户建模，并利用答案质量加权识别专家；S. Yarosh 等[10]基于专家的社会关系、自身专业知识等信息，构建“任务-主题”交叉场景，借助 SmallBlue Find 系统从推荐用户列表中选择专家；S. Ghosh 等[11]挖掘并分析了 Twitter 用户列表的元数据信息，利用 Cognos 系统查找主题专家。

(4)对用户的权威度、声望、参与度等进行建模或排序，从而达到识别专家的目的。如 D. R. Liu 等[12]通过主题偏好、声望和权威度的线性组合对用户建模：主题偏好由专家概况与目标问题的文本相似度算出，声望依据用户的历史答题数与最佳答案数，权威度由链接分析算法求得；L. Hong 等[13]依据问题主题对用户声誉进行建模，将概率潜在语义分析嵌入用户的声誉建模中，利用 PageRank 算法，进行专家发现；林鸿飞等[14]提出一种基于用户类别参与度的专家发现方法，利用 PageRank 和 HITS 计算了用户在每一个类别的专家得分及参与类别的参与度得分，帮助识别社区回答中的专家用户。

综上，学者们利用不同的测量指标及技术方法开展了专家发现的研究，但大多利用文本内容的相似性或辅助某一局部特征进行研究，在数据的模拟训练时大多采用单一的技术方法来论证模型的有效性，且研究的目的在于找到专家用户，多侧重技术方法，对理论部分的涉及也较少。本研究的目的不仅在于找到专家，而且要寻找回答问题可能性大的专家，基于社会资本理论和动机理论，从多个特征开展研究，从多个机器学习模型中寻找适合本模型的技术方法，一定程度上丰富了此前关于专家发现的研究。

2.2 知识共享

知识共享是指拥有知识的人将知识以某种形式表达并通过媒介分享的行为[15]。社会化问答社区用户回答问题这一行为属于知识共享。学术界对用户知识共享行为的研究主要是理论研究，迄今为止，研究成果已十分丰富。本文选择比较成熟的动机理论和社会资本理论作为研究的理论基础，对其相关研究进行梳理。

动机理论认为人们的行为由动机导向，动机是知识共享的必要前提[16]，学者们认为，虚拟社区中用户的动机可分为两大类，即内部动机(如个人兴趣[17]、个人乐于助人、利他主义与渴望被认同等因素[18-20])和外部动机(如声誉[18-21]、利益[17]、外部奖励[19-20,22]、获得有用信息及专业知识[23])，它们能让虚拟社区的用户有形或无形地获得一定的实质利益或是实现自我满足，显著地影响着用户知识共享的行为。因而，本研究结合研究对象的实际状况，从需求满足、利他主义的内部动机以及时间与利益的外部动机分析用户答题的可能性。

社会资本理论认为，社会资本、个人或社会网络所拥有的关系网络以及嵌入其中的资源集，强烈影响知识共享发生的程度，社会资本理论主要包括 3 个维度：结构维度、关系维度、认知维度[24]。赵玲等[25]、C. M. Chiu 等[26]、L. Zhao 等[27]、H. H. Chang 等[19]、H. F. Lin[28]认为社区互动关系，如信任、互惠等影响用户在虚拟社区的归属感(成员感)，进而影响用户知识共享的活跃度。B. Van den Hooff 等[29]认为社区的信任、认同感以及用户个人的知识共享能力和意愿是影响知识共享的重要因素。本研究围绕社会资本理论的 3 个维度，主要从互惠、成员之间的共同语言以及用户的社交关系网络进行用户答题潜在可能性的分析。

3 社会化问答社区答题者发现特征指标抽取

3.1 社会化问答社区用户答题动因分析

3.1.1 动机理论视角

动机理论视角下用户在社会化问答社区答题行为动因的分析如下(见图 1)：

一是内部动机下用户的行为动因分析，即用户出于满足自身某些需求或出于利他主义动机进行答题。

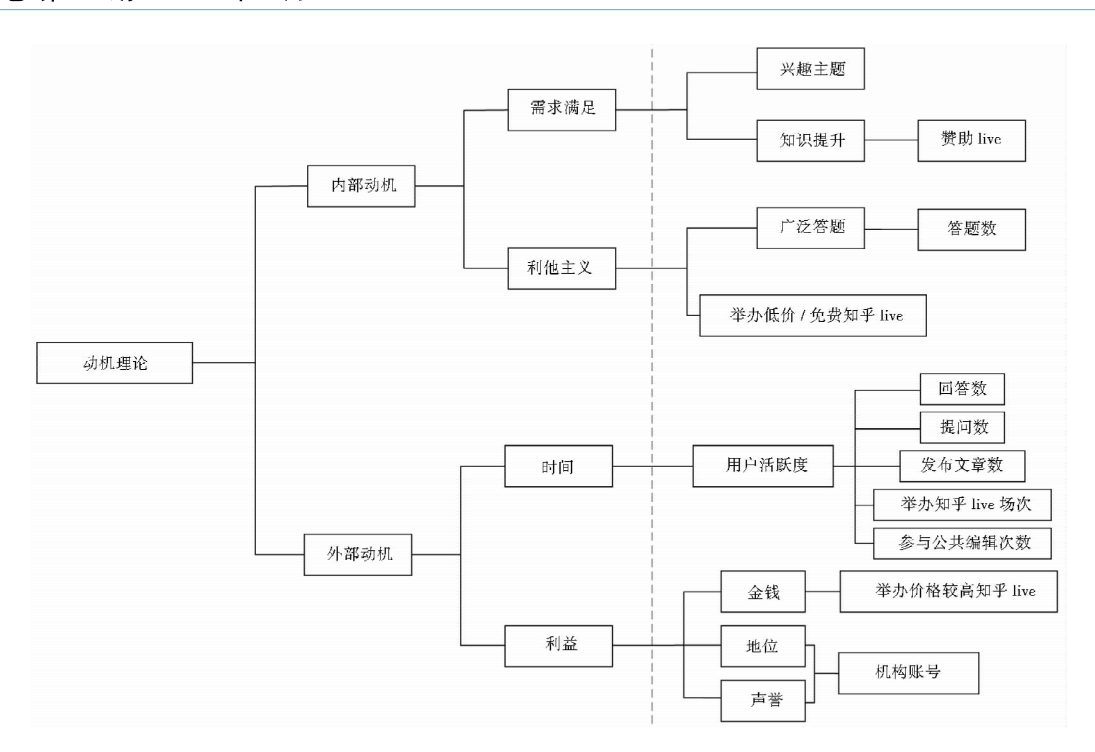


图 1 动机理论视角下用户答题动因分析

自身需求满足主要是指用户出于自我满足感，回答与其兴趣相关的问题；或是出于完善自身知识结构、提升自我的动机，而时常赞助 live 进行付费学习。（知乎 live 是知乎社区的实时问答，用户可实时参与或观看直播回放，在参与后用点亮星星的个数给出自己对本场 live 的质量、主题等方面的感受评价。用户也可以成为 live 的主讲人，通过语音、图片、视频或文字的形式实时分享自己的经验或见解，主讲人可以对举办的 live 进行免费或者付费设定，也可以选择 live 的主题。）利他主义动机下，用户则会广泛答题，或是举行超低价、免费的知乎 live 与社区其他用户交流信息，分享经验。

二是外部动机下用户的行为动因分析。主要是指用户具有在社区进行内容创作的时间，或是出于金钱等物质利益、提升声誉的动机做出某些行为。如外部动机下的时间因素是用户在社区活跃度水平高低的先决条件，金钱等物质利益及声誉也会诱使用户产生高质量的创作内容。金钱动机则表现在用户出于知识变现的目的举办付费 live，场次较多且价格不是很低。声誉动机则表现在机构认证用户及个人认证用户的创作行为习惯上（创作内容多提及与认证相关的信息）。

3.1.2 社会资本理论视角

社会资本理论视角下对用户在社会化问答社区答题行为动因的分析从关系维、认知维、结构维 3 方面展开，具体如下（见图 2）：

一是关系型社会资本，表现在社会化问答社区成员间的互惠关系上。互惠的表现方式有无形资源互惠和有形资源互惠两种。无形资源互惠，如用户 C 在内容创作完成后，创作内容被用户 D 浏览，用户 D 在浏览后为表达对内容的肯定进行了点赞或送出感谢等行为。有形资源互惠，如用户 D 举办非低价的付费知乎 live，在知识分享中实现知识变现，用户 F 通过付费获得用户 D 分享的知识，提升自己。

二是认知型社会资本，是指社区主体间的共同愿景和共同语言，表现为知乎用户因为某种兴趣、爱好、共同语言而聚集到同一类话题下交流信息，为丰富相关领域的专业知识而努力获取知识、分享知识。

三是结构型社会资本，是指社区用户在社交网络结构中所处的位置，主要与用户的粉丝数量、粉丝在社交网络中节点位置的重要性相关。

3.2 专家发现特征指标

根据以往学者对专家发现的研究，主要采用以下指标衡量社会化问答社区用户的专家身份：

一是用户可信度。社会化问答社区中用户既是受众，也是内容的发布者。平台内容信息的可信度与发布者密切相关，它代表着用户主观上对信息的信任程度，不仅仅是狭义上的“真”或“假”[30]。本文从用户的背景资料信息及用户在社区的交互行为两方面对用户的可信度进行评测。

二是用户专业性。问答社区中专家是回答过类似问题的用户[2]。本文从用户产生的历史答案的主题分

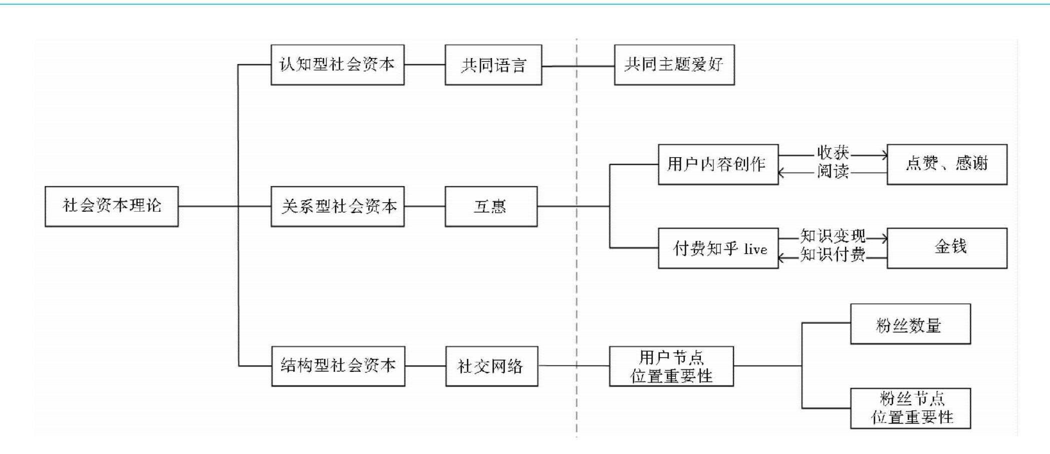


图2　社会资本理论视角下用户答题动因分析

布和内容质量两方面测量用户的专业性，包括内容是否详实、清晰、专业，即用户在回答时是否含有与问题相关的主题词，是否引用图表、链接进行内容的补充或辅助说明，回答是否仔细（答案长度）。

三是用户权威性。正所谓人以群分，用户在社会化问答社区也并非相互孤立的个体，他们总是倾向于跟在态度、兴趣、价值观、背景和人格上和自己相似的人进行在线社交[28]，形成一个个"圈子"，这些"圈子"是根据用户的兴趣倾向和所属知识领域而划分的社区结构[31]。用户在知乎社区进行问答及相关互动，久而久之呈现出"互粉""被粉""粉丝"3 种关注关系（见图3），用户成为"粉丝"或"被粉者"，由此形成庞大而清晰的用户社交关系网络。用户的权威值由用户之间形成的关系网络得出，可由 HITS 或 PageRank 计算[12]。

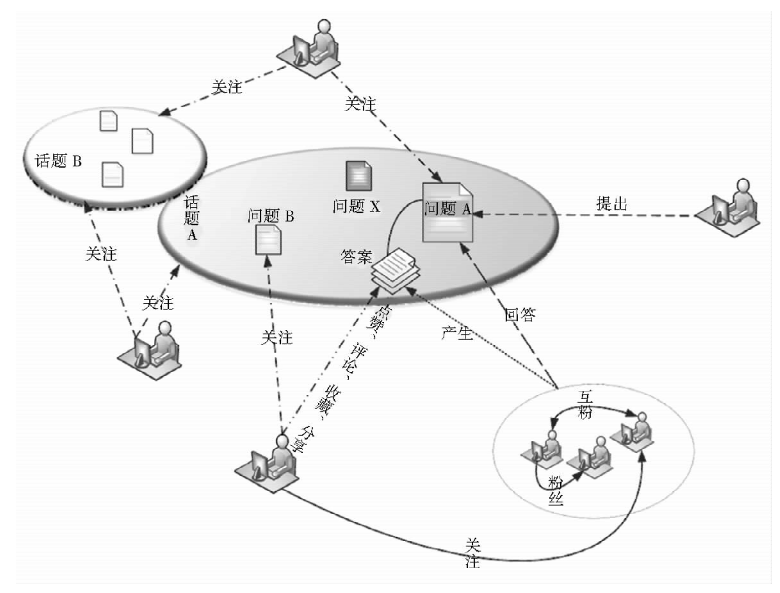


图3　知乎社区用户互动及用户关系

3.3　社会化问答社区答题者特征指标及指标测度

本节基于前文对用户答题动因分析及以往专家发现研究特征指标的提取，结合研究实例的状况提出本研究的特征指标（见图4）及研究指标的测量方法。

3.3.1　用户可信度

对知乎社区用户可信度的测量在一定程度上能够保证其产生答案的可信度。

用户背景资料是用户在注册、使用知乎时自己填写的简介资料，包括昵称、头像、性别、居住地、所在行业、职业经历、教育经历和一句话个人简介等内容。社会化问答社区中用户在注册时可选择实名注册，也可选择不实名注册。背景资料的完整程度能够在某种程度上反映出用户是否可信。本研究中用户资料完整度的统计采用一般的数学方法，即对用户的背景资料中包括用户性别、居住地、所在行业、职业经历、教育经历、个人简介字段不为空项的统计。此外，实名认证的个人用户或机构用户相比未认证用户、匿名用户而言，具有更高的可信度。

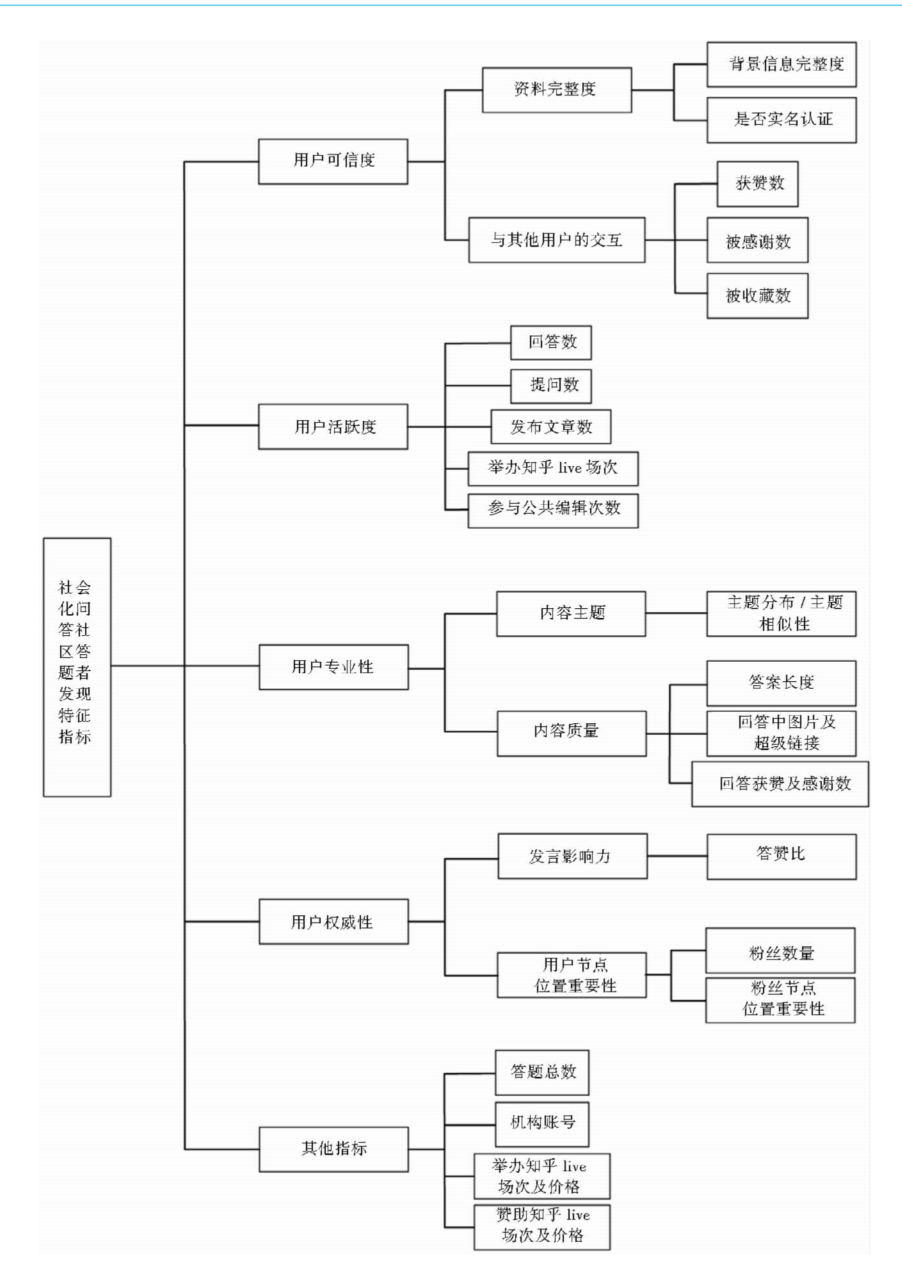


图 4　社会化问答社区答题者特征指标

用户在知乎社区与其他用户进行交互行为数据，能有效验证该用户是否为机器用户，是该用户可信度的又一体现。除了问答和关注关系，用户在知乎社区的交互行为主要通过用户产生的所有内容的获赞数、用户收到其他用户的感谢数、用户产生的内容创作被其他用户的收藏数 3 个指标共同衡量。实验中借助 TOPSIS 法对用户在社区的交互行为的相关数据进行降维的归一化处理，获取相应特征值，在尽可能减少原指标包含信息的损失的同时使得数据集更加易用、数据结果更易于理解，便于后续实验数据的处理。

3.3.2　用户活跃度

活跃度水平较高的用户更有可能为新问题给出自己的答案。用户在知乎社区的活跃度受众多因素的影响，如时间、个人兴趣、社区声誉及金钱等因素。本文基于用户在 2018 年一整年产生的历史回答数量、提问数量、发布的文章数量、举办的 live 场次数量，以及在社区中参与公共编辑的次数来衡量用户的活跃度水平。其中，用户在知乎社区参与的公共编辑是指用户在平台添加问题、为问题添加话题标签、移除问题、为问题移除话题标签、对问题进行补充说明等编辑行为。用户活跃度特征值的获取采用 TOPSIS 法。

3.3.3　用户专业性

在知乎社区中，大多数用户并未明确指出自己的兴趣主题或专业所在领域，对主题分布的描述有利于明确用户的主题兴趣，当用户主题与问题主题的相似

度较高时，用户更可能出于一种需求满足的内部动机进行答题。因此，为寻找专业的答题者使提问者得到满意回复，在面对一个新问题时，需要将新问题的话题关键词与用户的兴趣关键词进行匹配计算，主要包括：①用户个人简介信息与新问题之间的相似性。②用户主题与新问题的主题相似性。即将两者的信息点进行向量化，并计算其余弦距离。研究借助自然语言处理技术将用户回答的内容文本进行特征向量化，利用LDA（Latent Dirichlet Allocation）主题模型来抽取文本信息的关键话题特征。同时将新问题利用LDA主题模型转化为等维度的问题向量。其中，新问题与用户的话题匹配的余弦相似度为：

$$\mathrm{Sim}(\vec{H},\vec{V})=\frac{\vec{H},\vec{V}}{\|H\|\cdot\|V\|}\qquad 公式(1)$$

其中，H 和 V 是两个 n 维向量，H 是 $[H_1,H_2,H_3,\cdots,H_n]$，V 是 $[V_1,V_2,V_3,\cdots,V_n]$，余弦值越接近1，表明两个向量 H 和 V 越相似。

（1）tf-idf。处理文本信息时，需要对文本进行中文分词。本文将每位用户的个人简介或回答视作一个文档，所有用户文件为语料库，由此计算每个文档在语料库中的tf-idf。在处理过程中，利用Jieba对所有文本信息进行分词处理，同时要过滤无效的非常用词、标点、特殊符号等，接着计算每个中文词组的tf-idf，整体处理流程如图5所示：

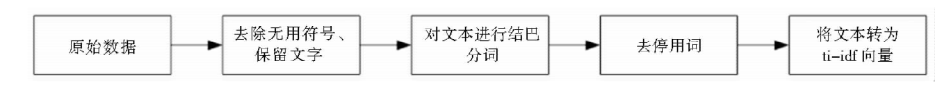


图5　中文文本预处理流程

（2）LDA主题模型。LDA是文档主题生成模型，又名三层贝叶斯概率模型，包含词语、主题和文档3层结构，其中，文档到主题、主题到词语均服从多项式分布。LDA能够抽取到大规模文档或语料库中潜在的主题信息。它采用词袋方法，将一篇文档看作是词频向量的构成，使复杂问题简单化。

3.3.4　用户权威性

本文从用户在社交网络中的重要性和用户产生的发言质量衡量知乎社区中用户的权威性。用户的权威性越高，吸引的粉丝就越多，其产生的答案被浏览、被转发、获赞、被收藏的几率就越大，其在社区发言所获得的影响力也越大，用户为提升影响力也更愿意答题，由此形成一个良性循环。

（1）用户发言影响力。研究依据用户回答对其他用户所产生的影响，包括获赞、转发、收藏来衡量用户在社区的发言影响力。知乎社区中高赞答案会优先显示，无形之中提升用户影响力，但知乎社区中单条答案的转发和收藏数量不予显示，故研究基于答案的获赞数来统计用户答案的获赞总数与用户产生的答案数量之比。

（2）PeopleRank算法。PeopleRank的计算原理跟PageRank相似，该算法将用户在社会化问答社区所形成的社交网络结构视作一个有向图，该图以社区的参与者——用户作为节点，用户之间的关注关系作为边（如图6中，用户A、B、C、D均为图的节点，若用户A关注用户B，则存在一条有向边 $A\rightarrow B$），用户A的粉丝越多，指向节点A的边越多，表明用户A在社交网络中的“圈子”规模越大，用户的重要性越高，即权威度越高。

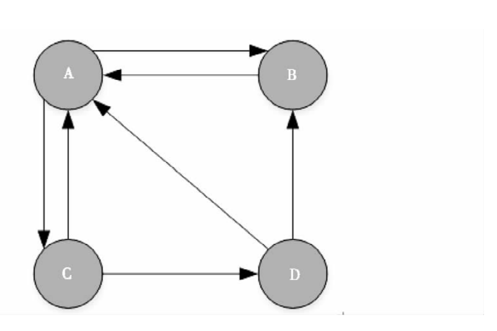


图6　用户知乎社交网有向图

对于任意用户A，其PeopleRank值为：

$$PR(A)=(1-d)+d\left(\frac{PR(p_i)}{C(p_i)}+\cdots+\frac{PR(p_j)}{C(p_j)}\right)\qquad 公式(2)$$

其中，p_i 代表用户，$C(p_i)$ 代表某个用户关注其他用户所形成的边的数量。d 是阻尼系数，代表用户间的关注关系可能改变用户权威度等级的概率。运算时，为每个用户赋予一个初始的 PR 值，通过算法不断迭代，直至 PR 值收敛稳定。

3.3.5　其他特征指标

研究中将无需计算可直接作为特征值的指标归于此类，主要包括用户是否为机构用户（机构用户会出于声誉及地位利益进行答题）、用户举办或赞助的知乎live场次及价格（用户出于提升自我的目的赞助知乎live进行学习，出于知识变现的目的举办较高价格的知乎live，出于利他主义动机多次举办免费或超低价知乎live），用户答题数（广泛答题是用户利他主义动机

的又一体现）主要采用一般的数学统计方法。

4　社会化问答社区答题者发现过程及算法

4.1　社会化问答社区答题者发现过程

本研究借助 Python 语言编写脚本从知乎网站采集数据，并对实验数据进行清洗和预处理，在此基础上，对实验数据进行特征提取，主要包括：①资料完整度；②用户活跃度；③用户在社区的交互行为；④用户历史创作特征；⑤用户在社交网络中的重要度；⑥问题话题与用户话题相似性。其中，问题话题与用户话题相似性由 tf-idf 及 LDA 主题模型提取社区用户的主题词及高频问题话题词向量之间的相似度求得。根据用户是否回答过该类主题的问题打上 0 或 1 标签（其中 0 代表用户未回答过该话题的问题，1 则代表回答过）。最后，将特征指标及标签值借助用户 ID 进行关联，将实验数据的 60% 作为训练数据，40% 作为测试数据，实验利用 3 种不同的机器学习模型（逻辑回归、随机森林、XGBoost）来构建准确率最优的二分类模型，并利用该模型计算不同用户对未回答的相关话题问题的答复概率，由此寻找合适的回答者，整体过程如图 7 所示：

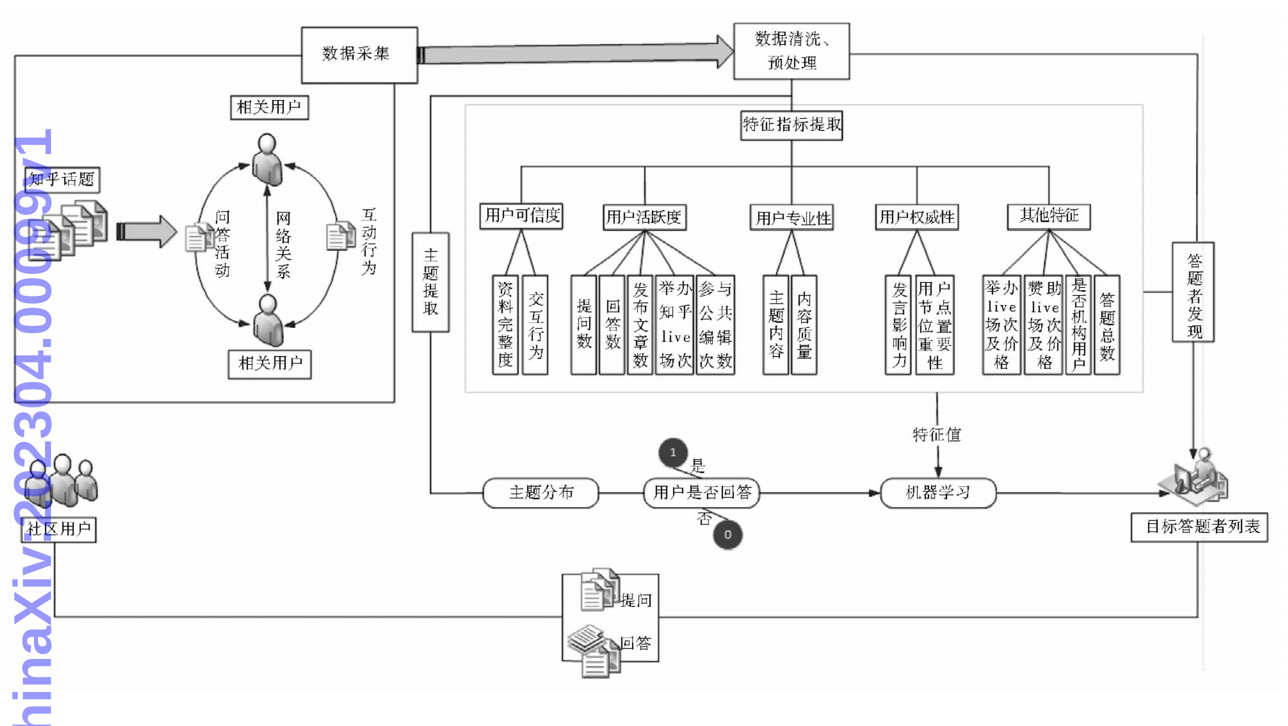


图 7　知乎社区潜在答题者发现过程

4.2　社会化问答社区答题者发现算法

技术模型是对文章理论指标的落地处理，基于前文的数据处理结果及相应特征值的提取工作，并分别选用逻辑回归、随机森林、XGBoost3 种模型对数据进行训练，对比得到最优的二分类模型作为预测模型。本节对社会化问答社区答题者发现模型的伪代码的描述见表 1。

5　实验构建及数据对比分析

5.1　数据集及预处理

5.1.1　实验数据集

本研究的实验数据来源于知乎社区，从医学话题领域入手，借助 Python 爬虫爬取用户及其相应数据，时间跨度为 2018 年 1 月 1 日至 2019 年 4 月 30 日，抓取的数据信息主要包括：①话题下用户的身份信息，包括用户 ID、所在行业、教育经历、个人简介等；②知乎社区对用户的认证信息，包括用户是否为机构用户、是否为知乎认证的话题优秀回答者；③话题下用户历史问答数据，包括用户提出的问题、用户已答问题、已答问题标签、用户产生的答案的数量及内容、答案被赞数、答案被感谢数；④话题下用户撰写的文章，包括标题、内容、标签、文章的赞同数；⑤话题下用户举办和赞助的知乎 live，包括 live 的题目、标签、价格、星级；⑥话题下用户的关注者及用户粉丝的 ID。

数据采集结果存放入 Access 数据库中，以用户在知乎注册时所形成的身份 ID 完成对各数据表的关联，数据集含 318 名用户的个人信息及提问 844 条，问答数据 65 352 条，关注数据 31 243 条，粉丝数据 276 379 条。

表 1　社会化问答社区答题者发现算法

```
Input:
用户数 i＝1,2,…,n(n＝306)
问题数 q

Output:每个用户对不同话题的输出概率
话题为 s,有 k 种,分别记为"医疗""游戏""科技""电影""食品""生活""就业""教育""亲密关系""收入"
For i＝1:q:
  对于每个问题,选择最恰当的话题标签
Return {"问题":"话题"}

For i＝1:n:
  1)每个用户计算其背景资料完整度
  2)每个用户计算其活跃度 AL 值
  3)每个用户在社区交互行为值
  4)每个用户是否未认证用户及 live 特征
  5)每个用户在社区社交网络中 PR 值
  6)每个用户创作内容特征值

  For j＝1:q:
  对于每个用户回答不同的话题,计算其相似度
  对于每个用户针对的具体话题,是否有过相应的历史回答,进行标签化{0 或 1}
Return 用户话题行为特征及相关标签数据

选择 3 类机器学习算法,利用不同的机器学习训练
For i in (LR,Random Forest,XGboost):
  F1,accuracy＝Comput(featuredata,label)
Model ＝ FindBestModel 找到准确率最佳的模型

得分预测:
Score 为用户特征数据与特定问题的推荐概率
```

5.1.2　数据预处理

对原始数据进行预处理时，考虑到匿名用户的数据信息在多数表格中出现缺省值(如背景资料的相关数据、提问数据等)，故研究剔除 12 位匿名用户及其产生单条问答信息，得到有效用户 306 位，与之对应的历史回答共 65 251 条，用户的关注数据共计 31 243 条，粉丝数据共计 276 379 条。

本次实验数据的预处理工作主要包含两个方面：①利用 SQL 语句对对各表中的数据进行去重。②利用 Python 语言清洗用户创作的文本内容，包括去除 html 标签、文本内容的分词及去停用词。借助 Python 的分词组件 Jieba 以及哈尔滨工业大学停用词表，并结合实验数据的实际情况添加了部分停用词、特殊符号及文字表情、序号以及一些常见的网络用语。

5.2　数据表征提取及分析

对实验数据表征的提取和分析不仅使数据更接近其背后代表的本质含义，也为后文数据标签的确定及进一步的数据分析和方法验证奠定基础。

5.2.1　用户可信度

(1)背景资料完整度。社会化问答社区中用户在注册时可选择实名注册，也可选择不实名注册。显然，实名注册用户更加可信，本研究中用户资料完整度的统计采用一般的数学方法，即对用户的背景资料(包括用户性别、居住地、所在行业、职业经历、教育经历、个人简介)6 个字段不为空项的统计，同时标注用户是否为认证用户。统计结果显示，大多数用户的资料完整度在 40% 以下，少数用户资料完整度在 60% 以上。在知乎社区进行实名认证的用户仅有 8 人，占比 2.6%。

(2)与其他用户的交互行为。用户与其他用户交互行为特征值由 TOPSIS 法依据用户在知乎社区的获赞数、被收藏数、被感谢数求得(见表 2)。首先确定该次评价指标均为极大型指标，接着对评价指标数据进行归一化处理，进而找出 3 个衡量指标的最优和最劣值，即 Z^+ 和 Z^-，最终计算出各评价对象(用户)与最优或最劣值间的距离 D^+ 和 D^-。根据 D^+ 和 D^-，得出各评价对象与最优值的接近程度(C 值)来表示用户的活跃度水平。

表 2　用户交互行为原始数据(部分)

项	获赞数	被收藏数	被感谢数
评价对象 1	45	306	40
评价对象 2	31	125	180
评价对象 3	58	394	38
…	…	…	…
评价对象 305	146	1 446	75
评价对象 306	5	21	2

以评价对象 1 为例，依据 TOPSIS 法对研究中的相关数据(见表 2)进行归一化处理得到：

$$\beta_{11}=\frac{45}{\sqrt{45^2+31^2+58^2+\cdots+146+5^2}}\approx 0.000\ 37$$

$$\beta_{12}=\frac{36}{\sqrt{306^2+125^2+394^2+\cdots+1\ 466^2+21^2}}\approx 0.000\ 26$$

$$\beta_{13}=\frac{40}{\sqrt{30^2+180^2+38^2+\cdots+75^2+2^2}}\approx 0.000\ 17$$

由此计算出所有的 β 值，进而得出 3 个衡量指标的最优和最劣值，即 Z^+ 和 Z^-，

$Z^+=\max\{\beta_{11},\beta_{21},\cdots,\beta_{3061}\}$，$\max\{\beta_{12},\beta_{22},\beta_{3062}\}$，$\max\{\beta_{13},\beta_{23},\cdots,\beta_{3063}\})=(0.957\ 8,0.983\ 5,0.725\ 8)$

$Z^-=\min\{\beta_{11},\beta_{21},\cdots,\beta_{3061}\}$，$\min\{\beta_{12},\beta_{22},\beta_{3062}\}$，$\min\{\beta_{13},\beta_{23},\cdots,\beta_{3063}\})=(0,0,0)$

进而得出：

$D_1^+=\sqrt{\sum_{i=1}^{3}(Z_1^+-\beta_{13})^2}\approx 1.552\ 389$，$D_1^-=$

$\sqrt{\sum_{i=1}^{3}(Z_1^- - \beta_{13})^2} \approx 0.000\ 487$，$C_1 = \frac{D_1^-}{D_1^+ + D_1^-} \approx 0.000\ 31$

因此，评价对象 1 的交互行为特征值为 0.000 3。类似地，求出所有评价对象的活跃度特征值，实验结果见表 3（此处仅展示 10 名评价对象相关数据）。

表 3　TOPSIS 评价用户行为计算结果

项	D +	D －	C
评价对象 1	1.552 388 985	0.000 486 841	0.000 313 509
评价对象 2	1.552 275 781	0.000 826 734	0.000 532 312
评价对象 3	1.552 278 958	0.000 608 915	0.000 392 118
评价对象 4	1.552 863 329	3.439 86E-06	2.215 17E-06
评价对象 5	1.552 865 508	0	0
评价对象 6	1.516 068 496	0.037 389 797	0.024 068 748
评价对象 7	1.552 846 975	2.528 11E-05	1.628 02E-05
评价对象 8	1.552 837 438	2.848 44E-05	1.834 31E-05
评价对象 9	1.552 864 419	1.719 93E-06	1.107 58E-06
评价对象 10	1.552 755 714	0.000 118 034	7.600 98E-05

5.2.2　用户活跃度

活跃度水平高的用户具有时间上的答题动机，更可能在知乎社区分享自己的见解和经验。用户在社区的活跃度主要通过用户产生的回答数、提问数、文章数、举办的 live 场次以及参与公共编辑次数 5 个指标衡量。同样地，该次评价指标均为极大型指标，依据 TOPSIS 方法，求得用户的活跃度水平（C 值）见表 4（仅展示 10 名用户）：

表 4　TOPSIS 评价用户活跃度计算结果

项	D +	D －	C
评价对象 1	1.500 005 496	0.034 339 233	0.022 380 390
评价对象 2	1.504 152 748	0.026 162 029	0.017 095 848
评价对象 3	1.500 101 431	0.044 342 443	0.028 710 945
评价对象 4	1.486 501 490	0.041 618 921	0.027 235 368
评价对象 5	1.443 895 068	0.107 015 983	0.069 002 012
评价对象 6	1.420 039 124	0.185 684 301	0.115 639 031
评价对象 7	1.472 803 374	0.086 876 242	0.055 701 338
评价对象 8	1.514 231 261	0.002 090 914	0.001 378 938
评价对象 9	1.419 561 581	0.195 188 857	0.120 878 652
评价对象 10	1.513 819 702	0.002 289 890	0.001 510 372

5.2.3　用户专业性

知乎社区中用户的专业性主要由用户发布的内容（包括内容主题和内容质量）来衡量。

（1）用户的兴趣话题及其与问题话题的相似性。研究借助 LDA 主题模型从每个用户发布的历史答案中提取 10 个主题，每个主题含 8 个主题词。同时按照话题热度统计用户在知乎社区提出的问题的话题标签（用户在知乎社区提出问题会有相应的话题标签），经统计，将实验数据中问题的话题按照热度大致分为 10 个话题，即“医疗”“游戏”“科技”“电影”“食品”“生活”“就业”“教育”“亲密关系”“收入”。实验中，将 LDA 主题模型获取的用户话题主题词与问题主题词构建主题词典，将问题话题词与用户话题词在词袋模型中进行向量表示，计算向量之间的相似性，作为问题话题与用户主题之间的相关性特征值。此外，针对用户是否回答过相关话题将数据标签化。如某一用户的主题词（见表 5）为“分子”“情感”“生活”“化学”“物理”“科技”“教育”“科研”，那么用户在问题主题词语中，与话题“医疗”“游戏”“电影”“食品”“就业”“收入”所对应的标签是“0”，与“科技”“生活”“教育”“亲密关系”话题所对应的标签是“1”，由此构建一张带标签的数据表。

表 5　某用户主题词

主题	主题词
Topic 0	作用力 分子 作用 原子 计算 物体 皮肤 纳米
Topic 1	伍佰 婚礼 一种 单独 孤独 爱情 记录 生命
Topic 2	感到 父母 喜欢 手机 经历 彩礼 感觉 事情
Topic 3	结构 泡沫 相关 化学 实验 过程 知识 研究
Topic 4	温度 蒸发 过程 速度 空气 水蒸气 沸腾 现象
Topic 5	表面 表面张力 液体 界面 作用 液滴 固体 重力
Topic 6	科学 技术 人类 世纪 发明 世界 欧洲 发现
Topic 7	老师 学生 学习 学校 大学 研究生 科研 导师
Topic 8	论文 科研 国内 工作 自由 一种 学术 期刊
Topic 9	物质 分子 压强 密度 体积 质量 增大 高度

（2）内容质量。对内容质量的描述包括内容是否详实、清晰、专业。涉及到的指标有答案中图片及超级链接的数量、回答获赞数、获得的感谢数及答案长度。

本研究通过 html 标签（主要是 < img > 及 < a href >）来统计用户回答时所引用的图片及超级链接数量，用户平均每条答案中对图片及超级链接的引用不足 0.3。用户回答获赞数及获感谢数在数据抓取时已有，无需另行统计。统计结果显示，绝大部分用户年内产生答案总数约为 100 条，少部分用户产生答案约 300 条，极少数用户产生答案非常之多，在 3 000 条以上，答案文本平均长度约 70 词，少部分用户回答较为详细，文本长度大于 100 词。

5.2.4　用户权威性

统计用户发布的答案总数及答案获赞总数后，定义答赞比为 γ，统计结果显示，大部分用户年内答案平均获赞数在 10 以下，占比约 83%，14.05% 的用户年内答案平均获赞数在 100 以上，500 以内，1% 的用户可高

潘梦雅，沈旺，代旺，等.社会化问答社区答题者发现研究[J].图书情报工作,2020,64(18):76－88.

达 1 000 以上，显然产生高赞答案的用户处于少数，如图 8 所示：

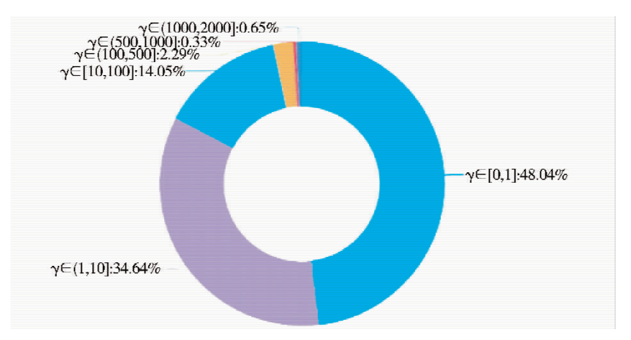


图 8　答赞比区间分布及用户占比

实验中，依据 31 243 条关注数据和 276 379 条粉丝数据，采用 PeopleRank 算法将用户作为图中的节点，用户之间的关注关系为图的边，由边的权重和边的数量计算每个用户的 PR 值，可视化结果见图 9。由于数据量过多，仅选取部分数据进行展示，某一节点的用户越多，说明该节点在社交网络中的重要性等级越高，用户权威性越高。

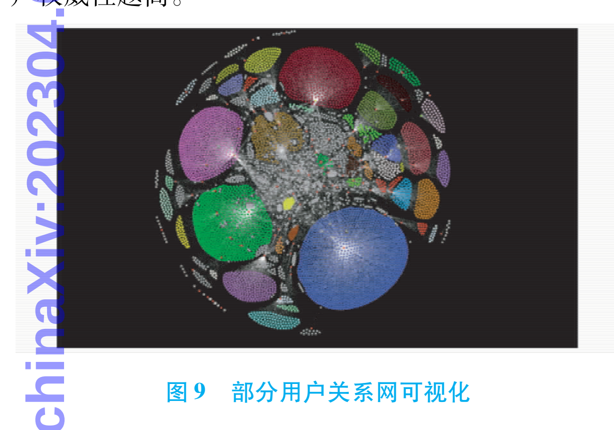


图 9　部分用户关系网可视化

5.2.5　其他特征值

实验数据中机构认证用户占比 0.6%，出于外部的利益动机，如声誉、地位、产品推广的目的，在答案的质量和数量上都非常可观。在知乎 live 方面，出于金钱利益，用户会设置较高的入场费，而那些乐于共享知识的用户，出于利他主义动机，会经常举办知乎 live，且大多免费或是价格极低（10 元以下），live 时长也较长（大于 2 小时）。少部分用户举办的知乎 live 均价达到 25 元以上，在分享知识的过程中实现了知识变现。多数用户（约 73.4%）出于提升自身知识水平的目的会赞助 live，个别用户场次较多，且费用达千元或以上。

至此，本研究的数据处理及数据表征的提取、分析工作全部完成。

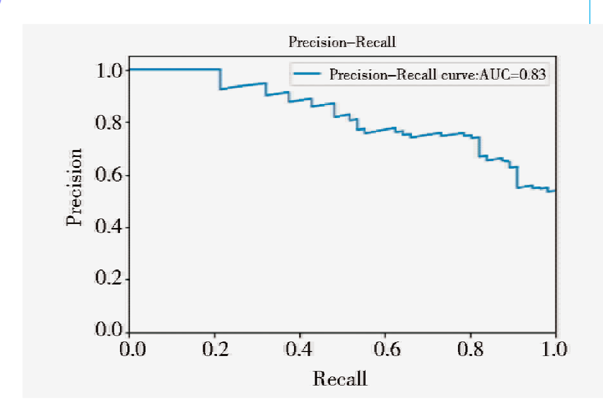
5.3　实验结果及对比分析

实验首先将样本数据集分成 60% 的训练数据与 40% 的测试数据，主要设计了利用 3 种不同的机器学习模型，来构建对用户回答相关话题问题的二分类预测模型对实验数据集进行训练，并在测试集中进行评估与对比。在测试集中的实验效果如表 6 所示：

表 6　3 种机器学习模型的实验结果对比

模型名称	准确率	精确率	召回率	f1	roc_auc
逻辑回归	0.598	0.570	0.791	0.661	0.666
随机森林	0.702	0.686	0.752	0.715	0.744
XGBoost	0.864	0.778	0.852	0.797	0.824

该二分类样本经过 3 种机器学习模型训练，得到的实验结果的最佳准确率为 86.4%，最佳的 f1 值为 79.7%。经过对比分析可知，利用 XGBoost 模型得到的预测效果最佳，其 PR 曲线及 ROC 曲线图如图 10 所示：



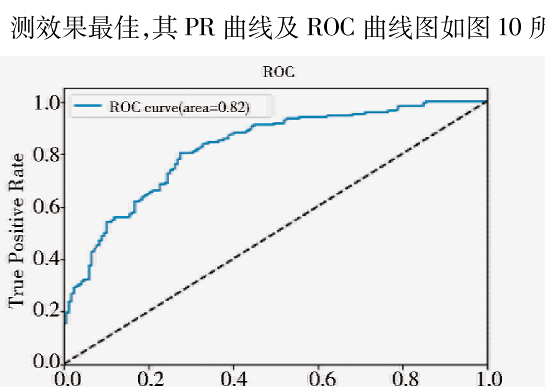


图 10　XGBoost 模型的 PR 曲线、ROC 曲线

由此，基于已训练好的模型，可预测用户与话题之间的推荐概率，用公式表示为：

$$Score = AL \cdot (\alpha_1 feature_1 + \alpha_2 feature_2 + \alpha_3 feature_3 + \cdots + \alpha_n feature_n) \quad \text{公式(3)}$$

其中，*AL* 代表用户活跃度，*feature* 为模型中的特征值。

本研究利用多种模型，并选择不同的模型超参数进行网格搜索，旨在计算出最佳的参数，利用训练好的

一组参数 $\alpha_1,\alpha_2,\cdots,\alpha_n$，可以预测用户与问题之间的回答概率，具体结果如表 7 所示：

表 7 部分用户回答问题的概率得分

$feature_1$	$feature_2$	$feature_3$	$feature_4$	$feature_5$	$feature_6$	score
0.35	0.24	0.02	0.29	0.24	0.73	0.83
0.84	0.23	0.01	0.29	0.23	0.69	0.12
1.00	0.24	0.05	0.29	0.24	1.00	0.92
0.66	0.17	0.21	0.22	0.17	0.34	0.16
0.88	0.29	0.08	0.35	0.29	0.77	0.27

通过模型，可以计算用户回答相关话题问题的概率得分。实验中，对上述结果的特征均进行标准化处理，测试的真实效果准确率大约为 79%。此外，为测试研究模型的优越性，本文将研究模型与 PageRank 算法、HITS 算法进行对比，结果如表 8 所示：

表 8 本模型与 PageRank 算法、HITS 算法比较

算法	准确率(%)	roc_auc
PageRank	76.3	0.738
HITS	69.1	0.621
本模型	86.4	0.824

由表 7 可见，利用本模型提取的指标结合众多的机器学习方法，生成的预测效果要略优于其他两种传统的模型。在考虑推荐的相关特征信息时，PageRank 算法和 HITS 算法没有关注用户本身在社会化问答社区的行为信息及用户主题相对于问题的偏好，从而导致其预测的准确率偏低。本文的研究模型一方面考虑了用户的活跃性、用户在社交网络中的专业性等诸多影响指标，还根据用户在社交网络的重要程度，来综合评价其推荐某个特定话题的子问题的概率。研究模型的预测效果也比较理想，能够对现实话题推荐起到一定的指导意义。

6 总结与展望

本研究以热门的社会化问答社区知乎社区为研究对象，从医学这一普通用户及具备专业知识答题用户都能参与的话题入手，进行时间跨度 2018 年一整年的实验数据的采集、清洗、预处理及特征分析。本研究基于社会资本理论及动机理论，构建相应的特征指标和研究模型，借助 Python 语言及相关算法将数据转换为模型所需的特征值，同时，依据用户是否进行过相应主题问题的回答给出 0 或 1 标签。实验时，选取实验数据的 60% 为训练数据，40% 为测试数据，运用逻辑回归模型、随机森林、XGBoost3 种常用的机器学习分类模型进行研究模型中数据的训练及预测，研究结果显示，该实验中 XGBoost 模型的准确率最高，能达到 86% 左右，拥有较好的实验效果。但研究也存在一些不足之处，如研究剔除了匿名用户的相关数据，但的确存在一些匿名用户，产生了切题度较高的答案，且论述也十分的详尽。此外，在用户答题动机的研究中，对用户主观的答题动机的测量较为客观，这些不足之处需要在未来的研究中给予完善，也为后续的研究指出新的思路和方向。

参考文献：

[1] LE L T, SHAH C. Retrieving people: identifying potential answerers in community question - answering[J]. Journal of the Association for Information Science and Technology, 2018, 69(10): 1246-1258.

[2] LIU X, CROFT W B, KOLL M, et al. Finding experts in community-based question-answering services[C]//Proceedings of the 14th ACM international conference on information and knowledge management. New York: Association for Computing Machinery, 2005: 315-316.

[3] WENG J, LIM E P, JIANG J, et al. Twitterrank: finding topic-sensitive influential twitterers[C]// Proceedings of the third ACM international conference on web search and data mining. New York: Association for Computing Machinery, 2010: 261-270.

[4] PAL A, COUNTS S. Identifying topical authorities in microblogs[C]// Proceedings of the fourth ACM international conference on Web search and data mining. New York: Association for Computing Machinery, 2011: 45-54.

[5] YAN Z, ZHOU J. Optimal answerer ranking for new questions in community question answering[J]. Information processing & management, 2015, 51(1): 163-178.

[6] CHENG X, ZHU S, CHEN G, et al. Exploiting user feedback for expert finding in community question answering[C]//Proceedings of the 2015 IEEE international conference on data mining workshop. Washington, D.C.: IEEE Computer Society, 2015: 295-302.

[7] SHEN J, SHEN W, FAN X, et al. Recommending experts in Q&A communities by weighted HITS algorithm[C]//2009 international forum on information technology and applications. New York: Institute of Electrical and Electronics Engineers, 2009: 151-154.

[8] PATIL S, LEE K. Detecting experts on quora: by their activity, quality of answers, linguistic characteristics and temporal behaviors[J]. Social network analysis and mining, 2016, 6(1):1-11.

[9] 龚凯乐，成颖. 基于"问题－用户"的网络问答社区专家发现方法研究[J]. 图书情报工作，2016，60(24)：115-121.

[10] YAROSH S, MATTHEWS T, ZHOU M. Asking the right person:

supportingexpertise selection in the enterprise[C]//Proceedings of the sigchi conference on human factors in computing systems. Austin: Association for Computing Machinery, 2012: 2247-2256.

[11] GHOSH S, SHARMA N, BENEVENUTO F, et al. Cognos: crowdsourcing search for topic experts in microblogs[C]//Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval. New York: Association for Computing Machinery, 2012: 575-590.

[12] LIU D R, CHEN Y H, KAO W C, et al. Integrating expert profile, reputation and link analysis for expert finding in question-answering websites [J]. Information processing & management, 2013, 49(1): 312-329.

[13] HONG L, YANG Z, DAVISON B D, et al. Incorporating participant reputation in community-driven question answering systems [C]//Proceedings of the 2009 international conference on computational science and engineering - volume 04. Washington, D. C.: IEEE Computer Society, 2009: 475-480.

[14] 林鸿飞,王健,熊大平,等.基于类别参与度的社区问答专家发现方法[J].计算机工程与设计,2014,35(1):333-338.

[15] SHARRATT M, USORO A. Understanding knowledge-sharing in online communities of practice[J]. Electronic journal on knowledge management, 2003, 1(2): 187-196.

[16] ARDICHVILI A. Learning and knowledge sharing in virtual communities of practice: motivators, barriers, and enablers[J]. Advances in developing human resources, 2008, 10(4): 541-554.

[17] RAZMERITA L, KIRCHNER K, NIELSEN P. What factors influence knowledge sharing in organizations? A social dilemma perspective of social media communication[J]. Journal of knowledge management, 2016, 20(6): 1225-1246.

[18] 黄维,赵鹏.虚拟社区用户知识共享行为影响因素研究[J].情报科学,2016,34(4):68-73,103.

[19] CHANG H H, CHUANG S S. Social capital and individual motivations on knowledge sharing: participant involvement as a moderator [J]. Information & management, 2011, 48(1): 9-18.

[20] CHO H, CHEN M H, CHUNG S. Testing an integrative theoretical model of knowledge-sharing behavior in the context of Wikipedia [J]. Journal of the American Society for Information Science and Technology, 2010, 61(6): 1198-1212.

[21] ZHANG Y, FANG Y, WEI K K, et al. Exploring the role of psychological safety in promoting the intention to continue sharing knowledge in virtual communities[J]. International journal of information management, 2010, 30(5): 425-436.

[22] WIERTZ C, DE RUYTER K. Beyond the call of duty: why customers contribute to firm-hosted commercial online communities [J]. Organization studies, 2007, 28(3): 347-376.

[23] BUTLER B, SPROULL L, KIESLER S, et al. Community effort in online groups: who does the work and why[J]. Leadership at a distance: research in technologically supported work, 2002, 54 (1): 171-194.

[24] NAHAPIET J, GHOSHAL S. Social capital, intellectual capital, and the organizational advantage[J]. Academy of management review, 1998, 23(2): 242-266.

[25] 赵玲,鲁耀斌,邓朝华.基于社会资本理论的虚拟社区感研究[J].管理学报,2009,6(9):1169-1175.

[26] CHIU C M, CHENG H L, HUANG H Y, et al. Exploring individuals' subjective well-being and loyalty towards social network sites from the perspective of network externalities: the Facebook case [J]. International journal of information management, 2013, 33 (3): 539-552.

[27] ZHAO L, LU Y, WANG B, et al. Cultivating the sense of belonging and motivating user participation in virtual communities: A social capital perspective[J]. International journal of information management, 2012, 32(6): 574-588.

[28] LIN H F. Determinants of successful virtual communities: contributions from system characteristics and social factors[J]. Information & management, 2008, 45(8): 522-527.

[29] VAN DEN HOOFF B, ELVING W, MEEUWSEN J M, et al. Knowledge sharing in knowledge communities[C]// HUYSMAN M, WENGER E, WULF V. Communities and technologies · January 2003. Netherlands: Kluwer, B. V., 2003: 119-141.

[30] 蒋盛益,陈东沂,庞观松,等.微博信息可信度分析研究综述[J].图书情报工作,2013,57(12):136-142.

[31] YANG C, DING H, YANG J, et al. Research of microblog community detection based on clustering analysis[J]. Advances in information sciences and service sciences, 2013, 5(3): 25-31.

作者贡献说明：

潘梦雅：论文撰写及修改、数据处理及分析；

沈旺：研究思路设计、论文最终版本修订；

代旺：数据分析；

刘嘉宇：论文修订。

Social Question Answering Community Respondent Discovery Research

Pan Mengya Shen Wang Dai Wang Liu JiaYu
Management School of Jilin University, Changchun 130022

Abstract: [Purpose/significance] Identifing the professional answerers with high probality in the social Q&A community can shorten the waiting time for users who ask questions to get satisfactory answers, promote knowledge sharing among users, and contribute to the sustainable and healthy development of the social Q&A community. [Method/process] Based on the social capital theory and motivation theory, this paper analyzed the motivation of users' answering questions, combined the expert discovery research to propose measurement indicators, and built a research model, then took Zhihu as a research example, and used Python to extract the eigenvalues and label of experimental data. Three common machine learning classification models, logistic regression model, random forest model and XGBoost model were used for training and prediction. [Result/conclusion] Compared with PageRank and HITS algorithms, the effectiveness and superiority of the method proposed by this paper have been verified. And this paper has provided a certain reference for the topic research of similar platforms such as healthy community problem push, expert identification and recommendation models.

Keywords: social Q&A community expert finding social capital theory motivation theory machine learning

《图书情报工作》2020年选题指南

[编者按]本选题指南是根据本刊的定位、性质与发展需要,结合图情档学科前沿热点及当前与未来需要解决的重要问题,邀请本刊编委和青年编委为本刊策划定制,再经编辑部整理、修改和补充而形成的。这是本刊2020年度关注、报道的重点领域(包括但不限于这些选题),供作者选题和研究以及向本刊投稿时的参考和借鉴。

1. 中国特色图情档学科体系、学术体系、话语体系建设
2. 图情档一级学科建设与融合发展战略
3. 图书馆"十四五"规划编制的关键问题
4. 国家文献信息资源保障能力及其建设
5. 开放科学背景下信息资源建设问题
6. 全民阅读中图书馆的定位与担当
7. 图书馆空间服务的理论与实践
8. 嵌入式学科服务的绩效评价与管理
9. 公众科学、科学素养与泛信息素养
10. 图书馆服务本科教育的模式与能力
11. 图书馆文化传承与文化育人的理论与实践
12. 图书馆出版与出版服务
13. 新媒体时代图书馆科学传播的功能与实践
14. 图书馆营销推广的战略与策略研究
15. 图书馆泛合作研究的实践与理论
16. 国家区域发展战略下图书馆联盟建设与创新服务
17. 网络空间治理的情报学问题
18. 知识产权信息服务能力与效果评估
19. 信息分析中的新技术与新方法
20. 情报服务标准化与评价
21. 数字人文与数字学术的研究与实践
22. 人工智能在图情档中的应用
23. 图书馆智能服务与智慧服务
24. 开放数据生态中的元数据发展模式研究
25. 开放科学数据行为及其模型构建
26. 数据资源建设与数据馆员能力建设
27. 大数据时代信息组织与知识组织
28. 科学数据管理与服务
29. 学术成果监测与学科竞争力分析
30. 情报计算(计算情报)的理论与方法
31. 情报分析服务质量与效能评价
32. 情报研究与智库研究的关系
33. 科学与技术前沿分析理论与方法
34. 健康中国2030战略下的健康信息学
35. 人机交互行为及服务模式创新
36. 图情档在新型智库建设中的作用机制
37. 智能信息服务的理论和方法
38. 数字公共文化资源、服务与体系建设
39. 数据时代政务信息资源管理和开发利用
40. 数字档案馆生态系统治理策略
41. 档案数据治理理论与治理体系
42. 政府数据开放平台应用与评价
43. 社会记忆视角下档案信息资源整理、保护与开发
44. 民族文献遗产产业化开发与利用
45. 图情档学科教育模式与人才培养能力